Learning Uncertainties the Frequentist Way: Calibration and Correlation in High Energy Physics

Rikab Gambhir
The NSF AI Institute for Artificial Intelligence and Fundamental Interactions
Center for Theoretical Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
rikab@mit.edu

Benjamin Nachman
Physics Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA
Berkeley Institute for Data Science, University of California, Berkeley, CA 94720, USA
bpnachman@lbl.gov

Jesse Thaler
The NSF AI Institute for Artificial Intelligence and Fundamental Interactions
Center for Theoretical Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
jthaler@mit.edu

Abstract

In this paper, we present a machine learning framework for performing frequentist maximum likelihood inference with Gaussian uncertainty estimation, which also quantifies the mutual information between the unobservable and measured quantities. This framework uses the Donsker-Varadhan representation of the Kullback-Leibler divergence—parametrized with a novel Gaussian Ansatz—to enable a simultaneous extraction of the maximum likelihood values, uncertainties, and mutual information in a single training. We demonstrate our framework by extracting jet energy corrections and resolution factors from a simulation of the CMS detector at the Large Hadron Collider. By leveraging the high-dimensional feature space inside jets, we improve upon the nominal CMS jet resolution by upward of 15%.

1 Introduction

One of the most foundational tasks in high energy physics (HEP) is the inference of an unobservable quantity given a measured quantity, which is often referred to as *calibration*. There has been significant progress in utilizing Machine Learning (ML) methods for calibrating the energies of various objects, including photons [1], muons [2], single hadrons [3, 4, 5, 6, 7, 8], and sprays of hadrons (jets) [9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19] at colliders; kinematic reconstruction in deep inelastic scattering [20, 21]; and neutrino energies in a variety of experiments [22, 23, 24, 25, 26, 27].

Abstractly, the calibration task can be described as quantifying the relationship between two random variables $X \in \mathbb{R}^M$ and $Z \in \mathbb{R}^N$. Here, X is the measured quantity and Z is the unobservable ("latent") quantity.[1] While ML methods are effective even when M and N are large, most existing methods have the undesirable property of being prior dependent [28]. As a result, the calibration is not universal and caution must be taken when applying it to different event samples.

[1]Throughout this paper, upper case letters represent random variables and lower case letters represent realizations of those random variables.

Machine Learning and the Physical Sciences workshop, NeurIPS 2022.

Furthermore, quantifying the reconstruction resolution is relevant for a variety of purposes, including the computation of significance variables [29, 30] and background estimation [31, 32]. Various ML approaches for resolution determination have been recently studied for HEP [33, 34, 35, 36, 37, 38, 39], but they typically require additional training or model complexity.

In this paper, we introduce a simple ML framework for calibration that simultaneously estimates the following quantities:

1. A prior-independent maximum-likelihood calibration, $\hat{z}(x) = \text{argmax}_z\, p(x|z)$;
2. A Gaussian resolution around $\hat{z}(x)$, $\hat{\sigma}_z(x)$;
3. The log-likelihood ratio, $\log \frac{p(x|z)}{p(x)}$; and
4. The mutual information between X and Z, $I(X;Z)$.

To extract $\hat{z}(x)$ and $\hat{\sigma}_z(x)$ in a single training, we use a novel *Gaussian Ansatz*, extending the Mutual Information Neural Estimator (MINE) of [40], to parametrize the log-likelihood ratio. After describing the Gaussian Ansatz construction, we illustrate the above features in a case study involving jet reconstruction at the Large Hadron Collider (LHC).

2 Calibration and Correlation

The starting point for our calibration method is the concept of mutual information (MI), defined as:

$$I(X;Z) = \int \mathrm{d}x\, \mathrm{d}z\, p(x,z) \log \frac{p(x,z)}{p(x)\, p(z)}, \tag{1}$$

where p denotes the probability density of the respective random variable. This equation has the property that $I(X;Z) = 0$ if and only if X and Z are independent. Therefore, the MI quantifies the interdependence between X and Z, including nonlinear correlations.

The MI is a special case of the well-known Kullback-Leibler (KL) divergence, $D_{\text{KL}}(P_{XZ}||P_X \otimes P_Z)$, where P_{XZ} is the joint probability distribution of X and Z, and $P_X \otimes P_Z$ is the product of the marginals. The KL divergence can be cast in the Donsker-Varadhan representation (DVR) [41]:

$$I(X;Z) = -\inf_{T \in \mathcal{T}} \mathcal{L}_{\text{DVR}}[T] \tag{2}$$

$$\mathcal{L}_{\text{DVR}}[T] = -\Big(\mathbb{E}_{P_{XZ}}\left[T\right] - \log\left(\mathbb{E}_{P_X \otimes P_Z}\left[e^T\right]\right)\Big). \tag{3}$$

Given a finite dataset of (x,z) pairs, the expectations in Eq. (3) can be estimated from sample averages. To estimate the second term, one can simply shuffle the x's and z's, as done in [40]. Then, the DVR loss functional can be minimized using standard gradient descent over parameterized neural networks T. For sufficiently expressive networks T, the infimum in Eq. (2) will be saturated, so the minimum loss is an estimate of $-I(X;Z)$.[2] Taking the functional derivative of the DVR loss functional with respect to T, we see that the minimum of $\mathcal{L}[T]$ is obtained when:

$$T(x,z) = \log \frac{p(x|z)}{p(x)} + c\,, \tag{4}$$

where c is an unimportant constant.Therefore, we can use T to extract the log-likelihood $p(x|z)$. This requires, as per the universal approximation theorem for machine learning, that the space of neural networks $\mathcal{T}$ is sufficiently expressive, that there is enough training data, and that the gradient descent algorithm successfully finds the minimum of Eq. (3). [3] Given this, which we will assume going forward, we can then perform maximum likelihood inference given x, and assuming that the likelihood is approximately Gaussian, even obtain the covariance matrix representing the inference resolution:

$$\hat{z}(x) = \underset{z}{\text{argmax}}\; T(x,z), \qquad \left[\hat{\sigma}_z^2(x)\right]_{ij} = -\left[\frac{\partial^2 T(x,z)}{\partial z_i\, \partial z_j}\right]^{-1}\Bigg|_{z=\hat{z}}. \tag{5}$$

[2] Numerical and analytic studies [40, 42], as well as our own empirical studies, show that the DVR loss has better numerical convergence properties than similar losses.

[3] We note that these assumptions are common to *every* machine learning method for inference, even if they are not explicitly stated.

Crucially, this inference strategy for z is independent of the prior $p(z)$, which is a property desirable for calibration tasks. Unlike for standard regression [28], the learned estimate $\hat{z}$ does not depend on the distribution of z samples in the training set.

However, both the maximum likelihood estimate and local resolution in Eq. (5) are difficult to evaluate numerically. The learned T may be highly non-convex and the true maxima difficult to find using gradient descent. Additionally, second derivatives are numerically sensitive to the choice of activation function in the neural network, especially the commonly used ReLU activation.

In order to facilitate a numerical estimate of the maximum likelihood and local resolution, we introduce the following Gaussian Ansatz parametrization for T:

$$T(x,z) = A(x) + \big(z - B(x)\big) \cdot D(x) + \frac{1}{2}\big(z - B(x)\big)^T \cdot C(x,z) \cdot \big(z - B(x)\big)\,, \tag{6}$$

where $A : \mathbb{R}^N \to \mathbb{R}$, $B : \mathbb{R}^N \to \mathbb{R}^M$, $C : \mathbb{R}^N \times \mathbb{R}^M \to \text{Sym}(M, \mathbb{R})$, and $D : \mathbb{R}^N \to \mathbb{R}^M$ are each neural networks. Unlike a Gaussian likelihood, the Gaussian Ansatz is highly expressive, and is in fact a universal function approximator. Specifically, any function $f(x, z)$ that admits a Taylor expansion in z around $B(x)$ can be expanded in this form. The functions $A(x)$, $D(x)$, and $C(x)$ capture the zeroth, first, and second (or higher) order dependencies of f on z, respectively.

The Gaussian Ansatz enables an elegant strategy to extract Eq. (5). Since the optimal $T(x, z)$ is bounded from above, we can take $D(x)$ to be everywhere zero without loss of expressivity.[4] In this case, T will achieve critical values at $z = B(x)$. Moreover, if $C(x, B(x)) < 0$, then these critical values will yield (local) likelihood maxima and resolution estimates:

$$\hat{z}(x) = B(x), \qquad \hat{\sigma}_z^2(x) = -\big[C(x, B(x))\big]^{-1}. \tag{7}$$

Moreover, the (negative) loss of the Gaussian Ansatz with respect to the functional in Eq. (3) will be a lower bound for the mutual information $I(X; Z)$, which is saturated in the asymptotic limit. The Gaussian Ansatz is therefore capable of estimating the maximum likelihood inferred value of z given x, the local resolution on that inference, and the mutual information between X and Z, all at once, with no additional postprocessing.

3 Case Study: Jet Energy Calibration

We now demonstrate the Gaussian Ansatz on a collider physics task: determining jet energy corrections (JECs) and resolutions (JERs) [43]. Jets are collimated sprays of particles that are produced ubiquitously in high-energy collisions. One does not have access to the "true" jet energy, however, because its constituent particles are filtered through a complicated and nonlinear detector response. This is an inherently prior-independent task, as it would be undesirable for energy corrections to depend on how often those energies appeared in the calibration set.This is an inherently prior-independent task, as it would be undesirable for energy corrections to depend on how often those energies appeared in the calibration set.

Assuming one has a good detector model (which one must assume anyways for any calibration method), though, one can *generate* truth-level quantities (GEN, corresponding to Z) and then *simulate* the detector response (SIM, corresponding to X). The JEC and JER factors are then defined such that the inferred jet momenta and resolution are:

$$\hat{p}_T \equiv \text{JEC} \times p_{T,\text{SIM}} \approx p_{T,\text{GEN}}, \qquad \hat{\sigma}_{p_T} = \text{JER} \times p_{T,\text{SIM}}, \tag{8}$$

where p_T is the transverse momentum of the jet.

We use the same 2011 CMS Open Simulation [44] samples as in [45], which are based on dijets generated in PYTHIA 6 [46] with a GEANT4-based [47] simulation of the CMS detector, in the MIT Open Data (MOD) HDF5 format [48]. Each SIM event consists of a list of particle flow candidates (PFCs), which are the reconstructed four-momentum and particle identification (PID)

[4] In practice, we find it convenient to start the training with non-zero $D(x)$ to aid the convergence of the model, and then numerically force $D \to 0$ through an increasing L_1 regularization. This helps the model achieve a global, rather than local, minimum. In our jet calibration studies, we find that this significantly improves model convergence.

Table 1: Jet Energy Calibrations

Model	Mean $\hat{p}_T$ [GeV]	Mean $\hat{\sigma}_{p_T}$ [GeV]	$I(X;Z)$
DNN	698 ± 37.7	35.7 ± 2.1	1.23
EFN	695 ± 37.3	32.6 ± 2.3	1.26
PFN	697 ± 36.9	32.5 ± 2.5	1.27
PFN-PID	695 ± 35.1	$\mathbf{30.8 \pm 3.6}$	**1.32**
CMS 2011	695 ± 38.4	36.9 ± 1.7	–

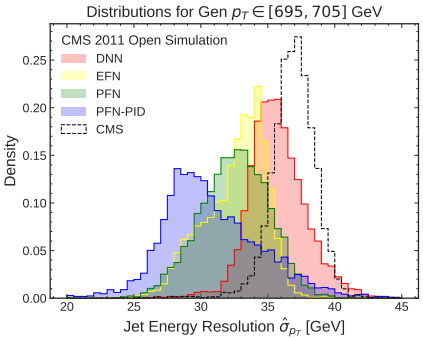

Figure 1: (Left) Gaussian Ansatz results for the four ML models, compared to the CMS 2011 baseline [43]. On a test dataset of GEN jets with $p_T \in [695, 705]$ GeV, we show the inferred $\hat{p}_T$, its resolution $\hat{\sigma}_{p_T}$, and the learned mutual information between $X = X_{\text{SIM}}$ and $Z = p_{T,\text{GEN}}$. (Right) Learned JER distribution for the four models, compared to the CMS 2011 baseline.

for each measured particle. The PFCs are clustered into anti-k_t jets with $R = 0.5$ [49, 50, 51]. For each jet, truth-level GEN jet information is also provided, as well as the CMS-prescribed JEC. CMS-prescribed JERs are estimated using [43].

We select jets whose GEN transverse momentum is in the range $p_T \in [500, 1000]$ GeV, whose GEN pseudorapidity satisfies $|\eta| < 2.4$, and that satisfy at least "medium" jet quality [52]. The latent variable of interest is $Z = p_{T,\text{GEN}}$, and the measured quantity $X = X_{\text{SIM}}$ is specified below. All momenta are divided by a fixed scale of 1000 GeV. In total, 5×10^6 jets are used for training.

We consider four different ML models, of increasing sophistication:

1. *DNN*: The input $X = (p_T, \eta, \phi)_{\text{SIM}}$ is the overall jet information, the same information used in the CMS calibration procedure in [43]. The functions A, B, C, and D are constructed as fully connected neural networks, with three hidden layers of size 64 and ReLU activations.
2. *EFN*: The input X consists of the entire set of PFC three-momenta. The functions A, B, C, and D are constructed as Energy Flow Networks (EFNs) [53]. For each EFN, the Φ and F functions (see [53]) consist of three hidden layers of respective sizes $(50, 50, 64)$ with ReLU activations. Since C is a function of both X and Z, the Z is appended as an input to the F function.
3. *PFN*: The same as the EFN, but all networks are Particle Flow Networks (PFNs) [54, 53] rather than EFNs.
4. *PFN-PID*: The same as the PFN model, but in addition to the 3-momenta of each PFC, the reconstructed PID is included as an input feature. We follow the PID labeling scheme of [53] for photons, charged hadrons, etc.

Each model is trained on a GPU cluster for 200 epochs using the ADAM optimizer [55], with a learning rate of $\alpha = 10^{-4}$ and a batch size of 2048. All model parameters are given an L_2 regularization loss of $\lambda_2 = 10^{-6}$. The D network is given an overall L_1 regularization loss of $\lambda_D = 10^{-3}$ to slowly force it to zero. Every 50 epochs, α is reduced by a factor of 5 and λ_D is increased by a factor of 10.

In Table 1, we show the results of the training in a narrow bin of $p_{T,\text{GEN}} \in [695, 705]$ GeV, though we note importantly that our results below are qualitatively similar across the entire p_T range, and that only a single bin was chosen for ease of interpretation and visualization. If our models yield unbiased estimators of the GEN p_T, then the inferred $\hat{p}_T$ distribution should be centered near 700 GeV, which it is for all models. We see indeed that the resolution improves with increasing model sophistication, as does the mutual information $I(X;Z)$, as expected. The PFN-PID model exhibits the best resolution, which is roughly 15% better on average than the CMS baseline.

In Fig. 1, we show the distribution of $\hat{\sigma}_{p_T}$ in the same $p_{T,\text{GEN}} \in [695, 705]$ GeV bin. As the model sophistication increases, the resolution increases (i.e. the $\hat{\sigma}_{p_T}$ shift downward). In principle, the resolution should never degrade by adding more information, but we do find a long right tail for the

PFN-PID model due to incomplete ML convergence.[5] We conclude that the measured PFC momenta, along with the PIDs, contain useful information for jet energy calibration that is lost when only considering the total jet momentum.

4 Conclusion

In this paper, we presented an extension of the MINE framework, the Gaussian Ansatz, capable of simultaneously performing frequentist inference, extracting Gaussian uncertainties, and quantifying mutual information between random variables. All of these tasks are performed in a single training, with no additional postprocessing. Using this framework, we can take advantage of the full jet particle information in the CMS Open Simulation to improve the measured jet resolution by approximately 15%. Studies by the ATLAS collaboration have used sequential calibration on a handful of observables to improve their resolution [56, 57, 58], and the Gaussian Ansatz may allow for further improvements by allowing for simultaneous calibrations of any number of features. We look forward to further developments in ML-based calibration and correlations methods in HEP and beyond.

Code and Data

The code for the general-use Gaussian Ansatz framework can be found here. The code and data for the jet energy calibration study, in particular, are available here.

Acknowledgments

We would like to thank Patrick Komiske for helpful discussions about EFNs and PFNs, Govert Nijs for helpful discussions on numerics and convergence, and Jennifer Roloff for helpful discussions about jet calibrations. We are grateful to Phiala Shanahan and Andrew Pochinsky for providing access to the Wombat cluster for some of the calculations undertaken in this work. RG and JT are supported by the National Science Foundation under Cooperative Agreement PHY-2019786 (The NSF AI Institute for Artificial Intelligence and Fundamental Interactions, http://iaifi.org/), and by the U.S. DOE Office of High Energy Physics under grant number DE-SC0012567. BN is supported by the U.S. Department of Energy (DOE), Office of Science under contract DE-AC02-05CH11231.

Broader Impact

While a few physics-inspired applications of a prior-independent calibration have been considered in this paper (and there are many more uses in physics, such the elimination of mass sculpting effects and adoption in many physics experiments), there are a variety of broader applications. Due to prior independence, the Gaussian Ansatz is a generic solution to the problem of imbalanced data sets, wherein certain data may be over- or under-sampled. This can often be the case in datasets where marginalized groups of people may be underrepresented in data. In addition, beyond its use in scientific contexts, the ability of the Gaussian Ansatz to do manifest uncertainty estimation is applicable to situations in which margins of error and safety in decision making are important, such as in self-driving cars.

References

[1] Albert M Sirunyan et al. Electron and photon reconstruction and identification with the CMS experiment at the CERN LHC. *JINST*, 16(05):P05014, 2021.

[2] Jan Kieseler, Giles C. Strong, Filippo Chiandotto, Tommaso Dorigo, and Lukas Layer. Calorimetric Measurement of Multi-TeV Muons via Deep Regression. 7 2021.

[5] We verified that the tail shrinks and the resolution improves with increasing training statistics, but we were limited by machine memory considerations.

[3] Dawit Belayneh et al. Calorimetry with deep learning: particle simulation and reconstruction for collider physics. *Eur. Phys. J. C*, 80(7):688, 2020.

[4] ATLAS Collaboration. Deep Learning for Pion Identification and Energy Calibration with the ATLAS Detector. *ATL-PHYS-PUB-2020-018*, 2020.

[5] N. Akchurin, C. Cowden, J. Damgov, A. Hussain, and S. Kunori. On the Use of Neural Networks for Energy Reconstruction in High-granularity Calorimeters. 7 2021.

[6] N. Akchurin, C. Cowden, J. Damgov, A. Hussain, and S. Kunori. Perspectives on the Calibration of CNN Energy Reconstruction in Highly Granular Calorimeters. 8 2021.

[7] L. Polson, L. Kurchaninov, and M. Lefebvre. Energy reconstruction in a liquid argon calorimeter cell using convolutional neural networks. 9 2021.

[8] Joosep Pata, Javier Duarte, Jean-Roch Vlimant, Maurizio Pierini, and Maria Spiropulu. MLPF: Efficient machine-learned particle-flow reconstruction using graph neural networks. 1 2021.

[9] ATLAS Collaboration. Generalized Numerical Inversion: A Neural Network Approach to Jet Calibration. *ATL-PHYS-PUB-2018-013*, 2018.

[10] ATLAS Collaboration. Simultaneous Jet Energy and Mass Calibrations with Neural Networks. *ATL-PHYS-PUB-2020-001*, 2020.

[11] Albert M Sirunyan et al. A Deep Neural Network for Simultaneous Estimation of b Jet Energy and Resolution. *Comput. Softw. Big Sci.*, 4(1):10, 2020.

[12] Rüdiger Haake and Constantin Loizides. Machine Learning based jet momentum reconstruction in heavy-ion collisions. *Phys. Rev. C*, 99(6):064904, 2019.

[13] Rüdiger Haake. Machine Learning based jet momentum reconstruction in Pb-Pb collisions measured with the ALICE detector. *PoS*, EPS-HEP2019:312, 2020.

[14] Pierre Baldi, Lukas Blecher, Anja Butter, Julian Collado, Jessica N. Howard, Fabian Keilbach, Tilman Plehn, Gregor Kasieczka, and Daniel Whiteson. How to GAN Higher Jet Resolution. 12 2020.

[15] Patrick T. Komiske, Eric M. Metodiev, Benjamin Nachman, and Matthew D. Schwartz. Pileup Mitigation with Machine Learning (PUMML). *JHEP*, 12:051, 2017.

[16] Convolutional Neural Networks with Event Images for Pileup Mitigation with the ATLAS Detector. Technical report, CERN, Geneva, Jul 2019.

[17] Benedikt Maier, Siddharth M. Narayanan, Gianfranco de Castro, Maxim Goncharov, Christoph Paus, and Matthias Schott. Pile-Up Mitigation using Attention. 7 2021.

[18] Gregor Kasieczka, Michel Luchmann, Florian Otterpohl, and Tilman Plehn. Per-Object Systematics using Deep-Learned Calibration. 3 2020.

[19] J. Arjona Martínez, Olmo Cerri, Maurizio Pierini, Maria Spiropulu, and Jean-Roch Vlimant. Pileup mitigation at the Large Hadron Collider with graph neural networks. *Eur. Phys. J. Plus*, 134(7):333, 2019.

[20] Markus Diefenthaler, Abduhhal Farhat, Andrii Verbytskyi, and Yuesheng Xu. Deeply Learning Deep Inelastic Scattering Kinematics. 8 2021.

[21] Miguel Arratia, Daniel Britzger, Owen Long, and Benjamin Nachman. Reconstructing the Kinematics of Deep Inelastic Scattering with Deep Learning. 10 2021.

[22] Junze Liu, Jordan Ott, Julian Collado, Benjamin Jargowsky, Wenjie Wu, Jianming Bian, and Pierre Baldi. Deep-Learning-Based Kinematic Reconstruction for DUNE. 12 2020.

[23] S. Delaquis et al. Deep Neural Networks for Energy and Position Reconstruction in EXO-200. *JINST*, 13(08):P08023, 2018.

[24] Pierre Baldi, Jianming Bian, Lars Hertel, and Lingge Li. Improved Energy Reconstruction in NOvA with Regression Convolutional Neural Networks. *Phys. Rev. D*, 99(1):012011, 2019.

[25] R. Abbasi et al. A Convolutional Neural Network based Cascade Reconstruction for the IceCube Neutrino Observatory. *JINST*, 16:P07041.

[26] M. G. Aartsen et al. Cosmic ray spectrum from 250 TeV to 10 PeV using IceTop. *Phys. Rev. D*, 102:122001, 2020.

[27] Kiara Carloni, Nicholas W. Kamp, Austin Schneider, and Janet M. Conrad. Convolutional Neural Networks for Shower Energy Prediction in Liquid Argon Time Projection Chambers. 10 2021.

[28] Rikab Gambhir, Benjamin Nachman, and Jesse Thaler. Bias and priors in machine learning calibrations for high energy physics, 2022.

[29] Albert M Sirunyan et al. Performance of missing transverse momentum reconstruction in proton-proton collisions at $\sqrt{s} = 13$ TeV using the CMS detector. *JINST*, 14(07):P07004, 2019.

[30] Benjamin Nachman and Christopher G. Lester. Significance Variables. *Phys. Rev. D*, 88(7):075013, 2013.

[31] Georges Aad et al. Search for squarks and gluinos with the ATLAS detector in final states with jets and missing transverse momentum using 4.7 fb^{-1} of $\sqrt{s} = 7$ TeV proton-proton collision data. *Phys. Rev. D*, 87(1):012008, 2013.

[32] Georges Aad et al. Search for new phenomena in events with an energetic jet and missing transverse momentum in pp collisions at $\sqrt{s}$ =13 TeV with the ATLAS detector. *Phys. Rev. D*, 103(11):112006, 2021.

[33] Albert M Sirunyan et al. A deep neural network for simultaneous estimation of b jet energy and resolution. 12 2019.

[34] Sanha Cheong, Aviv Cukierman, Benjamin Nachman, Murtaza Safdari, and Ariel Schwartzman. Parametrizing the Detector Response with Neural Networks. *JINST*, 15(01):P01030, 2020.

[35] Sven Bollweg, Manuel Haußmann, Gregor Kasieczka, Michel Luchmann, Tilman Plehn, and Jennifer Thompson. Deep-Learning Jets with Uncertainties and More. *SciPost Phys.*, 8(1):006, 2020.

[36] Marco Bellagente, Manuel Haußmann, Michel Luchmann, and Tilman Plehn. Understanding Event-Generation Networks via Uncertainties. 4 2021.

[37] Braden Kronheim, Michelle Kuchera, Harrison Prosper, and Alexander Karbo. Bayesian Neural Networks for Fast SUSY Predictions. 7 2020.

[38] Jack Y. Araz and Michael Spannowsky. Combine and Conquer: Event Reconstruction with Bayesian Ensemble Neural Networks. 2 2021.

[39] Braden Kronheim, Michelle P. Kuchera, Harrison B. Prosper, and Raghuram Ramanujan. Implicit Quantile Neural Networks for Jet Simulation and Correction. 11 2021.

[40] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm. Mine: Mutual information neural estimation, 2018.

[41] M. D. Donsker and S. R.S. Varadhan. Asymptotic evaluation of certain markov process expectations for large time—iii. *Communications on Pure and Applied Mathematics*, 29(4):389–461, July 1976. Copyright: Copyright 2016 Elsevier B.V., All rights reserved.

[42] Avraham Ruderman, Mark Reid, Dario Garcia-Garcia, and James Petterson. Tighter variational representations of f-divergences via restriction to probability measures, 2012.

[43] V. Khachatryan, A.M. Sirunyan, A. Tumasyan, W. Adam, E. Asilar, T. Bergauer, J. Brandstetter, E. Brondolin, M. Dragicevic, J. Erö, and et al. Jet energy scale and resolution in the cms experiment in pp collisions at 8 tev. *Journal of Instrumentation*, 12(02):P02014–P02014, Feb 2017.

[44] Cern open data portal.

[45] Patrick T. Komiske, Radha Mastandrea, Eric M. Metodiev, Preksha Naik, and Jesse Thaler. Exploring the space of jets with cms open data. *Physical Review D*, 101(3), Feb 2020.

[46] Torbjörn Sjöstrand, Stephen Mrenna, and Peter Skands. PYTHIA 6.4 physics and manual. *Journal of High Energy Physics*, 2006(05):026–026, may 2006.

[47] S. Agostinelli et al. Geant4—a simulation toolkit. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 506(3):250–303, 2003.

[48] Patrick Komiske, Radha Mastandrea, Eric Metodiev, Preksha Naik, and Jesse Thaler. CMS 2011A Open Data | Jet Primary Dataset | pT > 375 GeV | MOD HDF5 Format, August 2019.

[49] Matteo Cacciari and Gavin P. Salam. Dispelling the N^3 myth for the k_t jet-finder. *Phys. Lett.*, B641:57, 2006.

[50] Matteo Cacciari, Gavin P Salam, and Gregory Soyez. The anti-ktjet clustering algorithm. *Journal of High Energy Physics*, 2008(04):063–063, Apr 2008.

[51] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. FastJet User Manual. *Eur. Phys. J.*, C72:1896, 2012.

[52] Jet Performance in pp Collisions at 7 TeV. 2010.

[53] Patrick T. Komiske, Eric M. Metodiev, and Jesse Thaler. Energy flow networks: deep sets for particle jets. *Journal of High Energy Physics*, 2019(1), Jan 2019.

[54] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. Deep sets. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[55] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

[56] M. Aaboud et al. Jet energy scale measurements and their systematic uncertainties in proton-proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector. *Phys. Rev. D*, 96(7):072002, 2017.

[57] Georges Aad et al. Jet energy measurement and its systematic uncertainty in proton-proton collisions at $\sqrt{s} = 7$ TeV with the ATLAS detector. *Eur. Phys. J. C*, 75:17, 2015.

[58] Morad Aaboud et al. Determination of jet calibration and energy resolution in proton-proton collisions at $\sqrt{s}$ = 8 TeV using the ATLAS detector. *Eur. Phys. J. C*, 80(12):1104, 2020.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] We are careful to state that our theoretical results only hold in the well-trained limit, and even show an example where this is not true (PFN-PID).
   (c) Did you discuss any potential negative societal impacts of your work? [No] We are not aware of any negative societal impacts this work may cause.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes] The guidelines were consulted while writing this draft.
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] Yes - assumptions of sufficient expressivity and training convergence are repeatedly stated in 2

(b) Did you include complete proofs of all theoretical results? [Yes] All claims are justified in the text, or references to proofs provided.

3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] The code and data are available in the Code and Data section.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Hyperparameter details are specified in 3
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] The point of the paper is to include uncertainties inherent in the dataset, though no additional uncertainties are included.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No] We only provide
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See discussion of CMS Open Data and MOD in 3
   (b) Did you mention the license of the assets? [N/A] CMS Open Data does does not specify licensing
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No] The data is publicly available.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]